DOCUMENT RESUME

ED 218 313 TM 820 353

AUTHOR Green, Donald Ross
TITLE Matching Tests to Curricula and Test Equating.
PUB DATE Mar 82
NOTE 14p.; Paper presented at the Annual Meeting of the American Educational Research Association (66th, New York, NY, March 19-23, 1982).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Equated Scores; Grade Equivalent Scores; *School Districts; *Standardized Tests; Test Construction; *Testing Problems; Test Selection; Test Use; Test Validity
IDENTIFIERS Test Equivalence; *Test Publishers

ABSTRACT

Improved communication and cooperation between test publishers and school systems is essential for solving major test problems. Test-curriculum matching, test equating, and inappropriate use of grade equivalent scores are major problem areas. The degree to which a standardized test yields information relevant to what the system is trying to teach its students, determines which test is used. The basic skills measured in major standardized achievement batteries probably become unidimensional for most students beyond the primary grades. It has been proposed that school systems use custom tests suitable for their schools and then equate them to nationally-normed tests. However, tests not constructed to the same specifications cannot qualify as parallel tests and equating them is not feasible. Many tests are constructed and used in an inappropriate manner due to poor communication between test publishers and school systems. Grade equivalent scores are developed by most test publishers and most school systems use them. They serve the function of making scores more interpretable. (DWH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED218313

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

Matching Tests to Curricula and Test Equating

by Donald Ross Green

CTB/McGraw-Hill

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

D. R. Green

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 820 353

A Paper Presented to the Annual Meeting
of the
American Educational Research Association
New York City, March, 1982

In this paper I shall try to deal with two of the many areas where test publishers and school systems (federal, state, local) need to improve their communication and cooperation. I believe these topics are important and currently do not appear to be adequately understood by all concerned. The topics I have chosen involve the related matters of test-curriculum matching and test equating. Both have been relatively obscure until recently even though they involve very important aspects of the use of tests in schools.

First I shall briefly outline what I see to be the salient issues in these areas from the point of view of using tests in school systems and then point out how these technical issues have led to what I see as series of misunderstandings that only better communications can undo and/or prevent.

I don't expect to offer any concrete solution to the problems thus created partly because I believe more work must be done on the technical issues first. Rather I hope to increase awareness of the issues in the part of all concerned so that better communication between publishers and school systems can occur about these problems. A common understanding of the problem is a necessary albeit insufficient condition for good communication and cooperative solutions.

The first topic is the matter of the degree to which a standardized test given in a school system yields information relevant to what the system is trying to teach its students.

On the face of the matter it appears desirable for a school system to use tests which reflect what is being taught in its schools as closely as possible. Everybody, in fact, has been recommending that for decades. For example, an early book on tests and measurement stated ". . . any test must represent an extensive sampling of the materials of instruction. . ." (Ruch, 1929, p. 29). By the 1950s this idea was usually referred to as curricular validity and was considered an important element in test selection (e.g. Noll, 1957). In fact, the U. S. Office of Education in the shape of the Center for the Study of Evaluation spent much money, time and effort on developing a scheme for school systems to use in comparing their curricula to published achievement tests. In the course of so doing they fully documented the diversity of the major test batteries available in the early 1970s (Hoepfner, 1978). This phenomenon has been documented again in substantial detail by the Porter group (e.g. Freeman, Kuhs, Knappen & Porter, 1979, note 1) for fourth grade mathematics in several of these same tests. Almost certainly the currently available standardized achievement batteries differ from each other in content even more. Indeed, the publishers of these tests have insisted (and continue to insist) this is true, each trying to claim theirs is the best representative of the curricula in the nation's schools.

However, the meaning of these differences is much less than adequately explored. Some years ago it was demonstrated by the Anchor Test Study (Bianchini & Loret, 1972) that the reading portions of these tests correlated very highly in the middle grades, a result which has fostered the notion that the tests are all much the same. Of course high correlations do not prove identity but they do discourage claims of diversity. However, it is reasonable to claim that middle grade reading curricula are not as distinctive as they are at other grades and in other subjects. I am not aware of evidence that other portions of standard batteries are as closely related. Still it seems hard to believe that the fourth grade mathematics tests the Porter group have analyzed would not show similar correlations, especially since they are all rather old and were probably built to broader and less elaborate content specifications than the current editions of these widely used batteries.

It is relevant, therefore, to consider just how much difference variations in content within a single test might make in student scores. Are these tests all homogeneous as the high KR20s they report indicate? Porter and his colleagues have pointed out that while the tests are constructed to be empirically unidimensional, this does not establish that they measure the same thing over time or across populations differing in background (Porter, Schmidt, Floden & Freeman, 1978, note 2).

I suspect that in the early stages of learning, items measuring achievement in computation are often somewhat specific and relatively independent of other items in the set. Therefore, variations in content specifics can make a perceptible difference in outcomes. But general mathematical skills do develop, thank goodness, since there is just not enough time in the world to teach it all. By fourth grade most of what most pupils can do in arithmetic computation is substantially predictable from any reasonable subset of the general domain. It is obvious that students who can multiply numbers with several digits can also add them. It would be surprising if these same students couldn't subtract and it is at least probable they can do a fair amount of division.

In other words, the basic skills measured in the tests being discussed probably do become unidimensional for most students somewhere beyond the primary grades. To the degree this occurs the kinds of content differences between tests found in these areas may not be critical for valid measurement of growth in the area. Given a fair degree of homogeneity it becomes reasonable to let a few items represent each general area. It is the aggregation of mildly related items that is the basis of reliable measurement (Green, 1978). This is not to say that a test which precisely fits what was taught won't show so called "better"results than one that is broader and fits less well. But as Cronbach, (1971) among

others, has pointed out a test can fit the curriculum entirely too well for the results to support an adequate evaluation.

This argument suggests that in the primary grades the specific differences in the content of the tests is potentially more important than they would be later on. Fortunately the relevant domains are relatively small at these beginning levels so that reasonable matches of nationally standardized tests to local curricula are probably common for most systems. This conclusion is supported by the large growth shown on the lowest levels of these batteries by the students in most school systems. Bianchini (1978) has described a specific exception to this where curriculum fit made a large difference in grade 1 reading scores throughout a whole state. He argued further that such differences become much less important in reading around the fourth grade, asserting growth becomes relatively independent of the reading curriculum at that time.

One can further argue that "the curriculum" becomes more diverse and less clear cut as one goes up the grades within systems, schools and even classes. The curriculum of a school system is not a unitary matter. The Institute for Research on Teaching has shown that what teachers teach, the content of the textbooks used, the school system goals and objectives, and the tests given by the system all differ in

varying degrees (Porter, Schewille, Floden, Freeman, Knappen, Kuhs, and Schmidt, 1979, note 3). Furthermore, it is obvious that to some degree the curriculum is unique for many children. They are taught and/or perceive different parts of the curriculum.

In short, the matching of a test to the curriculum is always less than perfect. If the test is designed by the teacher of that class it is usually pretty close but that is one reason why the tests of a single teacher are not appropriate for the whole school, much less the school system. Similarly a test which reflects the peculiarities of a local curriculum is probably not appropriate for other systems, although this would presumably vary in degree depending on the factors previously discussed. Obviously no categorical assertion on this point is possible. But even if the assertion is not true, i.e., if the locally oriented test can function well in other systems one would still expect the students taking the curriculum it was built to fit to do better on it than others would. In other words, the test could not be used to indicate relative status of the home school system.

If the test does yield scores which show the effects of instruction in this school system more clearly than other tests can do it would seem to be most unlikely that this test measures the same things the other tests do. Thus it must lack generality and in particular it cannot stand in for a nationally normed test legitimately. This in turn means

that the only way one could establish national norms is to directly norm the local test. Now obviously no school system is going to go out and get national norms on its test. It may be less obvious but it is also true, given that the test is indeed measuring a unique curriculum, that such norms would have little meaning. The conclusion is that only local norms can be obtained and only local norms make sense anyway.

Now if the system is willing to settle for this -- and this is a perfectly reasonable thing to do -- the story ends. The system has its own batteries and evaluates itself via internally derived standards and goals. The relationship of these standards and goals to those found elsewhere is not known but perhaps nobody cares about that.

However, in some places people obviously do care and it is reasonable that they should -- for example, they may be concerned that the local goals may be too narrow or that the instructional program may generalize less well than some others or that the schools are inadequately staffed or equipped. In any case, the point is one can't really have one's cake and eat it too. It doesn't make any sense to get national norms on a local test.

Perhaps all this will seem obvious to some people but it has not been obvious to everybody. In fact, in recent years it has been widely proposed that school systems should use custom made tests uniquely suitable for these schools and then "equate" them to nationally normed tests.

This move has been fostered by the ease of equating when tests have been Rasch calibrated which has made both publishers and school systems consider this as practical and sensible. Furthermore, the U. S. Office of education sponsored an undertaking in Oregon designed to facilitate this (Holmes, 1980).

However, after looking at the results Holmes concluded that the enterprise did not work very well. It should not have been expected to do so. Tests not built to the same specifications can hardly be expected to measure the same things or qualify as parallel tests. Lacking this, equating is not feasible. Equivalent scores which have the same meaning across groups and situations ". . . may be derived only across test forms that are parallel, this is forms that measure, within acceptable limits, the same psychological function. . ." (Angoff, 1971, p. 539). Rasch calibration assumes unidimensionality and it was noted earlier that this condition does not always exist in achievement tests. If the calibration is defective, the equating must be. It is true that one can determine comparable scores for somewhat different tests and this may be useful but to quote Angoff again, ". . . scores scales are comparable only with respect to a specific group tested under specific conditions" (Angoff, 1971, p. 590).

These categorical statements do not mean that all this is crystal clear and easily understood anymore than is the matter of matching test content and curricula. For example,

there is room for argument about what constitutes "acceptable limits" of deviance from strict parallelism of tests being equated. Another example is the lack of clarity about unidimensionality and the model fit of items when using Rasch or other latent trait models. No fully agreed upon procedure for judging this exists and the consequences of violating this assumption are not fully known.

The result is that wishful thinking enhanced by poor communication and sometimes ignorance has generated a situation where both publishers and school systems, sometimes together and sometimes separately, end up building and using tests in inappropriate and unsound ways. Aside from calling for improved communication among all parties concerned, I have no solution to offer. However, it might help to close by briefly discussing a problem that has been with us for many years for much the same reasons. This latter problem is not obscure and does not entail any serious technical issues. I am referring to the use of grade equivalent scores.

"Everybody" knows or accedes to the proposition that these scores should be used only with caution or not at all, yet all major test publishers offer them or something very similar and almost all school systems accept the offer. Contrary to the assertions of some critics the reasons for this are not stupidity or ignorance or venality or even a desire to damage American education. The reasons as I see it are twofold.

The first is that, all their faults notwithstanding, grade equivalents really do serve the function of making scores more easily interpreted. They serve a need not well satisfied any other way. The second reason is that communication and cooperation between school systems and publishers are deficient. What we can do about this latter problem I don't know but I do think the first one can be solved. Suppose both the publishers and school systems agreed to only develop, report and use grade equivalent scores that fall within a two grade range from a child's placement. This would eliminate or sharply reduce those egregious errors which are commony cited as major faults of grade equivalent scores since it is the deviant scores that are most seriously misinterpreted. Furthermore, it would preclude finding mean GEs and making other inappropriate calculations with them; only order statistics could then be calculated and therefore their deficiencies as units matter little.

Coming to a general acceptance of this may be difficult but perhaps we can do so and then use the procedures to solve the problems I have discussed in this paper and others as they arise.

Reference Notes

1. Freeman, D., Kuhs, T., Knappen, L. & Porter, A. A closer look at standardized tests. East Lansing, Michigan: Institute for Research on Teaching, Michigan State University, 1979.

2. Porter, A. C., Schmidt, W. H., Floden, R. E., & Freeman D. J. Impact on what?: The importance of content covered. East Lansing, Michigan: Institute for Research on Teaching, Michigan State University, 1978.

3. Porter, A. C., Schwille, J. R., Floden, R. E., Freeman, D. J., Knappen, L. B., Kuhs, T. M., Schmidt, W. H. Teacher autonomy and the control of the content taught. (Research Series No. 24). East Lansing, Michigan: Institute for Research on Teaching, Michigan State University, 1979.

References

Angoff, W. H. Scales, norms and equivalent scores in Thorndike, R. L. Educational Measurement, American Council on Education, Washington, D. C., 1971.

Bianchini, J. C. Achievement tests in differentiated norms in Wargo, N. J. & Green, D. R. Achievement testing of disadvantaged and minority students for educational program evaluation. Monterey, CA.: CTB/McGraw-Hill, 1978.

Bianchini, J. C. & Loret, P. G. Anchor test study: Final Report vols. II, III, IV. Princeton, N. J.: Educational Testing Service, 1972.

Cronbach, L. J. Test validation in Thorndike, R. L. (Ed.) Educational Measurement. Washington, D. C.: American Council on Education, 1971.

Green, B. F. Jr. In defense of measurement. American Psychologist, 1978, 33, 664 - 670.

Holmes, S. E. ESEA Title I linking final project report. Salem, OR.: Oregon Department of Education, 1980.

Hoepfner, R. Achievement test selection for program evaluation in Wargo, N. J. & Green, D. R. Achievement testing of disadvantaged and minority students for educational program evaluation. Monterey, CA.: CTB/McGraw-Hill, 1978.

Noll, V. H. Introduction to Educational Measurement. Boston, Houghton - Mifflin Co., 1957.

Ruch, G. M. The objective or new-type examination. Scott, Foresman,: N. Y., 1929.